biology and life sciences forum

MDPI

Proceedings

Enhancing Grape Brix Prediction in Precision Viticulture: A Benchmarking Study of Predictive Models using Hyperspectral Proximal Sensors †

Maria Santos-Campos [1], Renan Tosin [1,2], Leandro Rodrigues [1,2], Igor Gonçalves [3], Catarina Barbosa[3,4], Rui Martins[2], Filipe Santos [2] and Mário Cunha [1,2]*

[1] Department of Geosciences, Environment and Spatial Planning, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, S/N, 4169-007, Porto-Portugal; up201805079@edu.fc.up.pt
[2] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, S/N, 4200-465, Porto-Portugal; mccunha@fc.up.pt
[3] ADVID - Associação para o Desenvolvimento da Viticultura Duriense, Edifício Centro de Excelência da Vinha e do Vinho Parque de Ciência e Tecnologia de Vila Real, Régia Douro Park, Portugal
[4] CoLAB Vines&Wines – National Collaborative Laboratory for the Portuguese Wine Sector, Edifício Centro de Excelência da Vinha e do Vinho Parque de Ciência e Tecnologia de Vila Real, Régia Douro Park, Portugal
* Correspondence: mccunha@fc.up.pt
† Presented at the title, place, and date.

Citation: Santos-Campos, M.; Tosin, R.; Rodrigues, L.; Gonçalves, I.; Barbosa, C.; Martins, R.; Santos, F.; Cunha, M. Enhancing Grape Brix Prediction in Precision Viticulture: A Benchmarking Study of Predic-tive Models using Hyperspectral Proximal Sensors. *Biol. Life Sci. Forum* 2023, 27, x. https://doi.org/10.3390/xxxxx.

Academic Editor: Gianni Bellocchi

Published: 08-11-2023

Abstract: Sustainable and efficient agricultural production is a growing priority in modern society. Viticulture, an important agricultural and food sector, also faces this challenge. Precision Viticulture (PV) has gained prominence as it aims to foster high-quality, efficient, and environmentally sustainable practices. The Soluble Solids Content (SSC) is essential for assessing grape ripeness and quality in the winemaking process. Conventional methods for determining SSC values (expressed in ºBrix) are invasive, expensive and labour-intensive, necessitating sample preparation, making large-scale analysis impractical. In response to these limitations, this study presents an innovative approach within the field of Precision Viticulture. It focuses on the non-invasive prediction of SSC using low-cost Proximal Hyperspectral Optical Sensors. These sensors rely on spectral reflectance measurements in the range of 340-850 nm. The study was conducted in a commercial vineyard in the Demarcated Douro Region, Cima-Corgo sub-region, Portugal, over six weeks during ripening. 169 grape berries from Touriga Nacional vines were analyzed under three irrigation regimes (no irrigation, 30% ETc, and 60% ETc). After organizing and preprocessing the data, machine learning algorithms, namely Partial Least Squares Regression (PLS), Random Forest (RF), and Generalized Linear Model (GLM), were applied to predict SSC values. These models' performance was thoroughly evaluated using cross-validation techniques. The performance of different models was evaluated showing significant differences, according to the metrics used (R2, RMSE and MAPE). The RF model demonstrated effectiveness and precision. A high R^2 value of 0.9312, coupled with low RMSE (0.9199 ºBrix) and MAPE (3.88%), signifies a strong fit to the data and accurate predictive capabilities. The results of this benchmarking study on predictive models of SSC provide valuable insights into the performance of various models, aiding winegrowers and winemakers in decision-making.

Keywords: grapes berries; machine learning; point-of-measurement; quality-gap; sugar content; *Vitis vinifera*

1. Introduction

The grapevine (*Vitis vinifera* L.) is a traditionally non-irrigated crop, but given the need to adapt to climate change on viticultural activity, several studies have shown that changes in the water status of the vine at critical phenological stages have a direct effect

on the composition and qualitative attributes of the grape, affecting vegetative growth, yield, the microclimate of the canopy and the metabolism of the fruit [1,2].

In the vineyard, careful management is essential to determine the harvest date and select grapes at optimum ripeness according to the desired characteristics. Some methods traditionally used to measure physiological parameters in the vineyard, such as Soluble Solids Content (SSC), expressed in ºBrix, are destructive, expensive, laborious as they require the preparation of samples and do not allow them to be spatialised from a PV perspective, making them less useful [3]. The real-time determination of processes related to abiotic stress and physiological processes related to fruit ripening (SSC, anthocyanins, carotenoids and organic acids) provides support for precision practices of great relevance to vineyards and wine [4]. The environmental conditions that most strongly influence °Brix include sunlight, temperature, and humidity. Irrigation timing also affects °Brix since reduced water availability during fruit development [5].

The common procedure used today by wine producers to evaluate the maturity of the grapes in a vineyard is using a refractometer [6]. Proximity hyperspectral optical technologies, such as Hyperspectral Optical Sensors (HOS), offer potential in the non-destructive and cost-effective assessment of grape ripeness in wine. However, high costs are still a challenge for most farmers. INESC TEC, in partnership with the *Faculdade de Ciências da Universidade do Porto* (FCUP), has been developing low-cost HOS and thus be able to have a cause-effect physiological adherence to the spectral data collected in vineyards. After collecting the spectral data, it was processed and analysed. Machine learning (ML), an area of AI, develops models to learn from data, improve performance by identifying complex patterns and use them in predictions. This strategy has wide applications, both in agronomic decisions regarding crop performance in a given environment and in supporting cultural practices [7]. Despite these promising technological advances, there are still factors limiting the full adoption of PV systems, particularly in validating this Proximal (HOS) under field conditions. These include shortcomings in terms of data acquisition, processing, and modelling to obtain useful information. Lack of high-throughput system for mapping spacio-temporal in vineyard to fill the quality-gap in the context of PV.

Monitoring grapevines over the ripening process in different hydric regimes, the main goal of this study is fto develop a predictive model of SSC based on proximal detection data. The specific goals include: i) testing the performance of low-cost sensor developed at INESC TEC, ii) to benchmark of ML models for SSC prediction, comparing the performance of each, using appropriate metrics, after applying pre-processing techniques to minimise undesirable effects, reducing data dimensionality and the matrix effect (spectral information on grape composition is characterised by multi-scale interference) and iii) analysing the performance of predictive SSC models in different hydric conditions.

2. Materials and Methods

2.1. Grape Sampling and data acquisition

This research, carried out in the Douro, was implemented at Quinta dos Aciprestes - Latitude 41.21° N; Longitude 7.43° W. The farm is located next to the river, at altitudes of between 100 and 350 metres, in the Douro Demarcated Region, sub-region of Cima-Corgo. The farm benefits from a Mediterranean climate, with two distinct seasons: the wet season from October to April and the dry season from May to September. The experimental design used was randomised blocks, in each block including 6 plants subjected to different irrigation treatments: No Irrigation (NR); 30% Crop Evapotranspiration (30% Etc) and 60% Crop Evapotranspiration (60% Etc). To obtain the most optimised model, it is crucial to evaluate the SCC of grapes from the initial ripeness stage until the ideal harvesting time. For each irrigation treatment, 4 grape berries were collected from each vine, in 2 rows and in 2 different locations in the row, over 6 weeks. Each week, one per vine was randomly selected from these berries, totalling 169 samples (Jul 28th: n=36, Aug 4th: n=36, Aug 11th: n=8, Aug 18th: n=18, Aug 25th: n=35, Sep 1st: n=36).

The SSC expressed in ºBrix, of the grapes was measured using an RHB-32ATC portable refractometer (Laxco Inc., Bothell, WD, USA) - destructive method. The refractometer measures from 0 to 32°Brix, with an accuracy of 0.20°Brix and a resolution of ±0.2°Brix. It was calibrated with a drop of distilled water and set to read 0 °Brix. The grapes previously measured by the spectroradiometer (next section) were carefully cut and pressed to use their juice for analysis with refractometer. For the spectral acquisition, the equipment acquires spectra covering the ultraviolet, visible and near-infrared zones, recording hyperspectral signatures between 340 nm and 850 nm of the electromagnetic spectrum [8]. This sensor has an LED light source (active sensor), which makes it possible to obtain spectra at night [9]. Hyperspectral point-of-measurement (HS-POM) measurements were taken by touching the berry to the light source, the power of the light source and the integration time were adjusted for optimal recording of the spectra within the linear quantification region and, finally, the grape spectra were stored. each spectrum was associated with the corresponding SSC reference measurements, resulting in the final dataset.

2.2. Modelation

The preprocessing data includes the filtering and reduction of spectral data. The hyperspectral data was filtered using the Savitzky-Golay digital filter, to smooth the data, reduce noise, and preserve the important characteristics of the signals. Different window sizes were evaluated, assessing the impact on the result. Due to the dimensionality of the data, Principal Component Analysis (PCA) was also conducted on the standardised data set. This means that the data was centred (by subtracting the mean of each variable) and scaled (dividing by the corresponding standard deviations) before the PCA analysis, ensuring that variables with different scales do not dominate the principal components (PCs) due to their magnitudes and mitigating the matrix effect.

Throughout this study, different ML methods were evaluated to predict the SSC. The ML methods considered were: Random Forest (RF); Generalised Linear Model (GLM) and Partial Least Squares (PLS) [10]. RF is a model that is tolerant of data noise, its performance is high in determining spectral reflectance measurements due to its low sensitivity to outliers [11]. For this reason, two tests were conducted using RF, with spectral data filtered without PCA selection and with PCA selection, adjusting hyperparameters by Random Search. The PLS model was evaluated using Leave-One-Out Cross-Validation (LOOCV) and selection of the number of components (ncomp) based on the criterion of the lowest RMSE value. Finally, the GLM model can be used to model a variety of distributions for the dependent variable, such as in the situation where the data does not follow a normal distribution. The model was trained with a Gaussian distribution with cross-validation for the selection of PCs and different hyperparameters were evaluated (distribution family, link functions) as well as other linear regression, Ridge and LASSO Regressions.

To assess the generalisation capacity of the different models evaluated, the data was divided into two different sets by random sampling without replacement: the training set (70% of the data) and the validation set (30%). To assess the performance of each model evaluated and select the most robust, the following metrics were used: Coefficient of determination R^2; root mean square error (RMSE); mean absolute percentage error (MAPE). Residual analysis: tests and analysis distribution of residual in the irrigation treatments.

3. Results

Analysing the average spectral curves per irrigation treatment (Figure 1), according to the absorption of the photosynthetic pigments, the reflectances of the irrigation treatments indicate that the NR treatment has the highest concentration of chlorophyll a (428 and 453 nm), chlorophyll b (642 and 661 nm) and carotenoids (400 and 500 nm) [12]. On the other hand, the 60% Etc treatment shows lower reflectances in the chlorophyll and xanthophyll range (540-580 nm) [13] and anthocyanins (550 nm) [14].

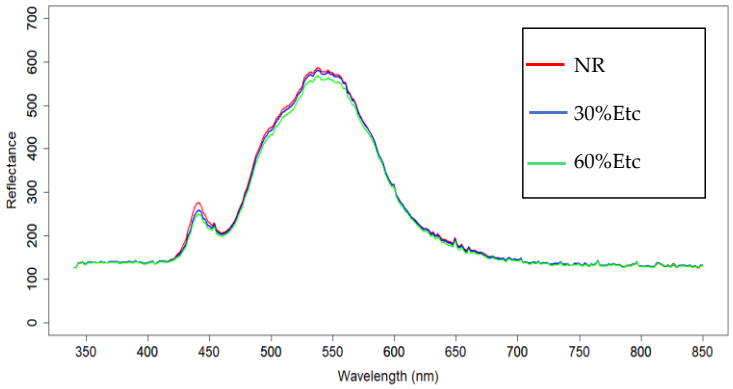

Figure 1. Spectral averages per irrigation treatment from July to September.

The SSC predictive capacities of the models evaluated were assessed on the training and the validation sets and are shown in Table 1. Comparing the two results obtained for each model evaluated allows the most robust to be selected and overfitting to be analysed.

Preprocessing: Although 5 components explain 99% of the total variance, 3 to 15 PCs were considered for training and validating the different ML models, according to the results of the cross-validation applied to assess their performance.

Table 1. Performance of the different ML models obtained in the training and validation sets for determining the SSC.

Models	Training Set			Validation Set		
	R^2	RMSE (ºBrix)	MAPE (%)	R^2	RMSE (ºBrix)	MAPE (%)
RF	0.9895	0.3988	1,49	0.9312	0.9199	3.88
PCA+RF	0.9427	0.9072	3.66	0.7134	1.8585	8.35
PCA+PLS	0.6427	2.0696	0.08	0.6382	2.0414	0.09
PCA+GLM	0.9991	0.1009	0.00	0.9990	0.1076	0.01

The PCA+GLM model shows excellent results in the training and validation sets, with extremely high R^2 values and low RMSE and MAPE. However, these extremely good results may show overfitting due to the high complexity [15]. Cross-validation stratified in relation to the different lambda values returned (data not showed) an RMSE of 3.45 ºBrix and a MAPE of 17.72 %, with relatively high cross-validation errors (MSE between 9.5 and 12.5). The RF model shows a slight drop in R^2 in the validation set compared to the training set. The PCA+RF model shows a significant drop in R^2 and an increase in RMSE and MAPE in the validation set compared to the RF model. The PCA+PLS model has a moderate performance in terms of R^2 on the validation set, but an extremely low RMSE and MAPE. Given these results, the RF model was the most effective.

The SSC was estimated using the training and test sets of TN samples collected over the period under analysis. Figure 2 shows the results of the SSC estimation on both sets, allowing us to compare the predicted SSC values with the observed ones and conclude on the model's performance in terms of its generalisation capacity. According to the metrics obtained in the validation set, the model explains 93.1 % of the variance in the data (R^2 = 93.1 %), with an average error of the predictions compared to the observed values of 0.920 (RMSE = 0.920 ºBrix) and an average accuracy of the predictions deviating 3.9 % from the real values (MAPE = 3.88 %).

Figure 3 presents the model's residuals plotted against the observed ºBrix.

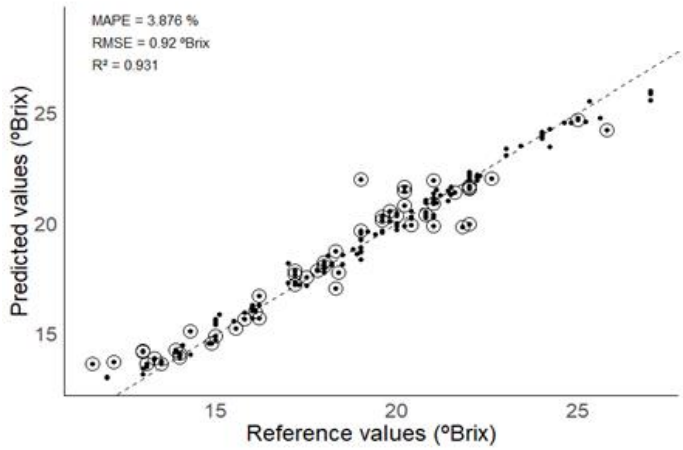

Figure 2. Results of the predictions from the valida-tion set with Touriga Nacional (TN) berry samples (dots circled) when applied to the Random Forest model trained with TN samples from the training set (dots).

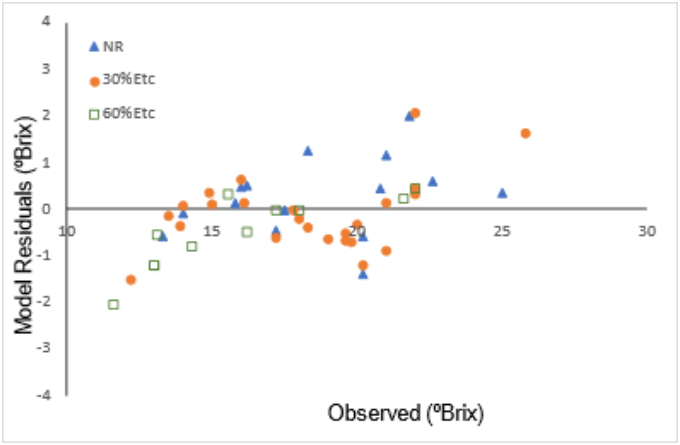

Figure 3. Distribution of model residuals (validation set) for different irrigation regimes.

The residuals spread on both sides of the "zero line" with no tendency of the residuals. The Shapiro-Wilk test results returned the W-Statistic value = 0.96607 and the p-value = 0.1512 (p_value>0.05). The ANOVA (F = 0.262; p>0.05) confirmed no significant differences between the irrigation treatments.

4. Discussion

In the spectral averages per irrigation treatment, 2 peaks of reflectance are visible. Reflectance in the 420-460 nm range is related to the absorption of chlorophyll a [12], which are green pigments involved in photosynthesis. However, the significant presence of these pigments in ripe grapes is unusual, as the plant directs energy towards the production of ripening-related compounds rather than chlorophyll. Therefore, it may be more related to carotenoids. Carotenoids are common in many fruits and vegetables, including grapes, as they are responsible for colours ranging from yellow to red. Visible reflectance in the 500-600 nm range may be related to anthocyanins, which are the pigments responsible for the red and purple colours in red grapes [14].

The application of low-cost hyperspectral optical sensors with Machine Learning models for precision viticulture presents a promising alternative to destructive and expensive conventional techniques. The developed RF model seems to be the safest choice in terms of overfitting, as it performs well in the validation set and has a moderate difference between the training and validation sets. The model's evaluation metrics show that

can explain approximately 98.95% of the variability in SSC values with high precision. The PCA+RF model shows a significant drop in R^2 and an increase in RMSE and MAPE in the validation set compared to the RF model. This suggests that dimensionality reduction with PCA led to a loss of valuable information and poorer performance, leading to overfitting, due to the loss of information during dimensionality reduction. The PCA+PLS model has a moderate performance in terms of R^2 in the validation set, but an extremely low RMSE and an almost zero MAPE. This may be indicative of a model with a high bias. Finally, the PCA+GLM model shows results suggesting overfitting, which occurs when an ML model overfits the training data, including noise, reducing its performance on new data sets. This means that the model fits the training data so well that it cannot generalise effectively to independent data, thus affecting the model's ability to predict accurately. Essentially, the model learns not only the actual structure of the data but also random fluctuations, which reduces its usefulness in real situations [15].

The RF model is the most robust choice, as it performs well in the validation set and there is slight difference between the training and validation sets. The residuals analysis confirmed the null hypothesis (H0) that the residuals follow a normal distribution, and their homoscedasticity is satisfied, i.e., the variance of the residuals does not vary significantly as the predicted values increase. The RF model proved to be effective and accurate. The high R^2 value (0.9312), the relatively low RMSE (0.9199 ºBrix) and MAPE (3.88%) indicate that the model is well adjusted to the data and can make accurate predictions. This model was tested on a dataset with high variability in SSC values. The results demonstrate that the irrigation treatments did not significantly impact the model's performance, which indicates the potential generalization of the model's results. These results are in line with other studies carried out in the Douro region, R^2 value (0.959) and RMSE (1.026 ºBrix) [16], as well as in other regions, namely the Mediterranean, R^2 value (0.83) and RMSE (1.99 ºBrix) with an RF model applied to the Syrah grape variety [17].

5. Conclusions

This work allowed benchmarking of SSC predictive ML models. Differences in the performance of 4 models tested on TN grape berries collected throughout the ripening period were demonstrated. The RF model is the most robust, not only because it is one of the models with the highest rate of explanation of the variation in SSC by the independent variables (R2 = 0.9312), but it is also the safest choice in terms of overfitting. The MAPE values suggest that the model can make good predictions on both the training data (1.49% MAPE) and the test data (3.9% MAPE). The different irrigation treatments and the ripening date did not have a significant impact on the predictive abilities of the model.

The potential demonstrated by some of the models justifies the investment in low-cost Hyperspectral Optical Sensors, such as Metbots. Evaluating the generalization capacity using different vintages could generate new studies related to the rapid and non-destructive assessment of the ripeness of wine grapes.

Funding: This research was funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots: High-Throughput Integrative Omic-Robots Platform for a Next Generation Physiology-based Precision Viticulture, with reference PTDC/ASP-HOR/1338/2021. https://www.fc.up.pt/omicbots/

Acknowledgments: Renan Tosin and Leandro Rodrigues acknowledge Fundação para a Ciência e Tecnologia (FCT) PhD research grants Ref. SFRH/BD/145182/2019 and SFRH/BD 2023.01424. Rui Martins acknowledges Fundação para a Ciência e Tecnologia (FCT) research contract grant (CEEIND/017801/2018). The authors thank the wine company Real Companhia Velha and Associação para o Desenvolvimento da Viticultura Duriense (ADVID) for the field work facilities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Acevedo-Opazo, C., S. Ortega-Farias, and S. Fuentes, Effects of grapevine (Vitis vinifera L.) water status on water consumption, vegetative growth and grape quality: An irrigation scheduling application to achieve regulated deficit irrigation. Agricultural Water Management, 2010. 97(7): p. 956-964.

2. Cunha, M. and C. Richter, The impact of climate change on the winegrape vineyards of the Portuguese Douro region. Climatic Change, 2016. 138(1): p. 239-251.

3. Tosin, R., et al., Assessing predawn leaf water potential based on hyperspectral data and pigment's concentration of Vitis vinifera L. in the Douro Wine Region. Scientia Horticulturae, 2021. 278.

4. Martins, R.C., et al., Unscrambling spectral interference and matrix effects in Vitis vinifera Vis-NIR spectroscopy: Towards analytical grade 'in vivo' sugars and acids quantification. Computers and Electronics in Agriculture, 2022. 194: p. 106710.

5. Kleinhenz, M.D. and N.R. Bumgarner, Using °Brix as an Indicator of Vegetable Quality: Linking Measured Values to Crop Management. Agriculture and Natural Resources, 2013.

6. Reid, M. and A. Kader, Postharvest technology of horticultural crops. Univ. Calif. Coop. Ext. Serv. Div. of Agr. and Natural Resources. Spec. Publ, 2002. 3311.

7. Sharma, A., et al., Machine Learning Applications for Precision Agriculture: A Comprehensive Review. 2021. 9: p. 4843-4873.

8. Martins, R., et al., Metbots: Metabolomics Robots for Precision Viticulture. 2019. p. 156-166.

9. Tosin, R., et al., Sensores proximais hiperespectrais e algoritmos computacionais para fenotipagem digital de parâmetros fisiológicos da videira. A revista da Associação Portuguesa de Horticultura. Nº 146.out 58: 24-26, 2022.

10. Pôças, I., et al., Toward a generalized predictive model of grapevine water status in Douro region from hyperspectral data. Agricultural and Forest Meteorology, 2020. 280: p. 107793.

11. Chea, C., et al., Optimal models under multiple resource types for Brix content prediction in sugarcane fields using machine learning. Remote Sensing Applications: Society and Environment, 2022. 26: p. 100718.

12. Lichtenthaler, H. and C. Buschmann, Chlorophylls and carotenoids: Measurements and characterization by UV-Vis spectroscopy. Food Analytical Chemistry: Pigments, Colorants, Flavors, Texture and Bioactive Food Components, 2005: p. 171-178.

13. Middleton, E., et al., Spectral Bioindicators of Photosynthetic Efficiency and Vegetation Stress. Hyperspectral Remote Sensing of Vegetation, 2011: p. 265-288.

14. Steele, M., et al., Nondestructive Estimation of Anthocyanin Content in Grapevine Leaves. American Journal of Enology and Viticulture, 2009. 60.

15. Montesinos-López, O., A. Montesinos, and J. Crossa, Overfitting, Model Tuning, and Evaluation of Prediction Performance. 2022. p. 109-139.

16. Gomes, V.M., et al. Determination of sugar content in whole Port Wine grape berries combining hyperspectral imaging with neural networks methodologies. in 2014 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES). 2014.

17. Kalopesa, E., et al., Estimation of Sugar Content in Wine Grapes via In Situ VNIR–SWIR Point Spectroscopy Using Explainable Artificial Intelligence Techniques. Sensors, 2023. 23(3): p. 1065.